Chapter 6

Using Observer Ratings to Assess Situation Awareness[1]

Herbert H. Bell
Air Force Research Laboratory
Human Effectiveness Directorate

Don R. Lyon
Link Simulation and Training

This chapter reviews the development and use of observational measures to assess *situation awareness* (SA) among fighter pilots. The chapter begins with a general description of air combat and SA. The next two sections summarize the general approach and the results of this effort. The chapter concludes with a discussion of the advantages and disadvantages of this approach and some general comments on the problems involved in measuring concepts that are as ill-defined as SA.

Uncertainty is characteristic of air combat. This uncertainty places enormous demands on the pilot's cognitive resources (Houck, Whitaker, & Kendall, 1993). The pilot must execute multiple tasks under extreme time pressure. At the same time, the pilot must deal with a variety of data sources, each of which may present only limited information about the current environment. For example, data regarding the location of enemy aircraft may come from on-aircraft systems that are controlled by the pilot. That data may also come from radio calls made by other members of the flight or an air weapons controller. Much of this data often corresponds to what the pilot already knows. Other pieces of data, however, provide new or conflicting information. Consequently, the pilot must filter, analyze, and interpret this data to estimate its timeliness and accuracy. The pilot must synthesize this data, assess the situation, and select a course of action.

[1]The views expressed in this chapter do not necessarily reflect those of the United States Air Force or the Department of Defense.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JAN 2000	**Final**	**00-01-1997 to 00-12-1999**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Using Observer Ratings to Assess Situation Awareness	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER **62205F**
6. AUTHOR(S) **Herbert Bell; Don Lyon**	5d. PROJECT NUMBER **1123**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory/RHA,Warfighter Readiness Research Division,6030 South Kent Street,Mesa,AZ,85212-6061**	8. PERFORMING ORGANIZATION REPORT NUMBER **AFRL; AFRL/RHA**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory/RHA, Warfighter Readiness Research Division, 6030 South Kent Street, Mesa, AZ, 85212-6061**	10. SPONSOR/MONITOR'S ACRONYM(S) **AFRL; AFRL/RHA**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AFRL-RH-AZ-BC-2000-0001**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Published as Chapter 6 In M.R. Endsley & D.J. Garland (Eds.). Situation awareness analysis and measurement (pp 129-146). Routledge USA

14. ABSTRACT
This chapter reviews the development and use of observational measures to assess situation awareness among fighter pilots. The chapter begins with a general description of air combat and situation awareness. The next two sections summarize the general approach and the results of this effort. The chapter concludes with a discussion of the advantages and disadvantages with this approach and some general comments on the problems involved in measuring concepts that are as ill-defined as situation awareness (SA).

15. SUBJECT TERMS
Situation awareness; Analysis; Measurement; Observer ratings; Situation assessment processes; Training programs; Human factors; System design

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**18**	

The necessity of making decisions based on the perception and interpretation of incomplete and sometimes conflicting data is one reason that training and experience are so important to the development of SA.

Endsley (1995a) views SA as an hierarchically organized construct with three interrelated levels. At its most basic level, SA involves the perception of current environmental information. At Level 1 SA, a pilot knows factual data such as the aircraft's energy state and the locations of other aircraft. The next higher level of SA reflects the pilot's interpretation of the current environmental data in terms of its immediate significance to the pilot's goals and objectives. An example of Level 2 SA is recognizing relative offensiveness or defensiveness during an air combat engagement. At the third and highest level of SA, the pilot not only recognizes and comprehends the current situation but also uses that information to anticipate future environmental states. At Level 3 SA, the pilot in an air combat engagement knows that both aircraft are currently neutral. In addition, the pilot knows that, based on the current flight paths and energy states, the enemy will soon have an offensive advantage.

In 1991, the Air Force Chief of Staff asked a series of questions about SA. These questions included: what is SA; can we measure SA; can we select individuals for pilot training based on their SA potential; and what impact does training have on SA. In response to these questions, Armstrong Laboratory (now the Air Force Research Laboratory) initiated a SA research program. This chapter summarizes our initial attempts to measure SA in operational fighter squadrons and in multiship air combat simulations.

Our initial efforts focused on three issues: the definition of SA; the degree to which pilots can reliably judge their fellow pilots in terms of SA; and whether there is a relation between such judgments and mission performance.

In response to the question, what is SA, the Air Staff provided a working definition that links SA to mission performance. This definition, written from the pilot's perspective, defines SA as "a pilot's continuous perception of self and aircraft in relation to the dynamic environment of flight, threats, and mission, and the ability to forecast, then execute tasks based on that perception" (Carroll, 1992, p. 5). Although there are a number of other definitions of SA available (e.g., Endsley, 1995a; Sarter & Woods, 1991; Tenney, Adams, Pew, Huggins, & Rogers, 1992), the Air Staff definition was used as the basis for our research efforts. This definition reflects the importance of SA in mission accomplishment, thus capturing the richness and complexity of the concept. It emphasizes perceiving what is important and then using that perception to guide the selection and execution of appropriate behaviors. Unfortunately, it is also very complex because it combines processes, tasks, and the linkages between them into a single construct. Consequently, it is very difficult to separate SA from the other aspects of skilled performance that determine combat proficiency.

MEASURING SA IN OPERATIONAL FIGHTER SQUADRONS

With the assistance of instructor pilots and other subject-matter experts (SMEs), Waag and Houck (1994) identified 31 behavioral elements of SA. The SMEs felt these elements reflected SA and were important to mission success. Table 6.1 lists these 31 elements and the 8 categories of mission performance the elements represent.

SA Instruments

The Air Force Research Laboratory developed 4 different instruments to measure SA in operational F-15C squadrons based on these 31 elements. The first instrument required respondents to provide their personal definition of SA. Using their personal definition of SA, each respondent then rated the importance of the 31 elements using a 6-point Likert scale.

The other three instruments, or SA Rating Scales (SARS), measured SA from three different perspectives: self, supervisory, and peer. The self-re-

TABLE 6.1
Categories and Elements of SA

General Traits	*Information Interpretation*
Discipline	Interpreting air-to-air radar
Decisiveness	Interpreting radar warning receiver
Tactical knowledge	Ability to use air weapons controller
Time-sharing ability	Integrating overall information
Reasoning ability	Radar sorting
Spatial ability	Analyzing engagement geometry
Flight management	Threat prioritization
Tactical Game Plan	*System Operation*
Developing plan	Radar
Executing plan	Tactical electronic warning system
Adjusting plan on-the-fly	Overall weapons system proficiency
Communication	*Tactical Employment—Beyond Visual Range*
Quality (brevity, accuracy, timeliness)	Targeting decisions
Ability to effectively use information	Fire-point selection
Tactical Employment—General	*Tactical Employment—Within Visual Range*
Assessing offensiveness/defensiveness	Maintain track of bogeys/friendlies
Lookout (radar, electronic, visual)	Threat evaluation
Defensive reaction (chaff, flares, maneuvering)	Weapons employment
Mutual support	

port SARS and supervisory SARS required the respondents to rate either themselves or their subordinates on each of the 31 items. Both SARS used a 6-point scale and the ratings were made relative to other F-15C pilots with whom the respondents had flown. The scale anchors were *acceptable* and *outstanding* because all the pilots were on flying status and were considered mission-ready by Air Force standards. The squadron commander, operations officer, assistant operations officer, weapons officer, and standardization-evaluation flight examiner completed the supervisory SARS on the pilots within their squadron. In addition, squadron flight commanders completed supervisory SARS on the pilots within their flight. The peer SARS required respondents to rate the other mission-ready pilots in the squadron on general fighter pilot ability and SA ability and then to rank order them on their SA ability. Both the peer and supervisory SARS allowed respondents to omit rating and ranking a particular pilot if they felt they did not have enough information to accurately judge that individual. All respondents completed the self-report and peer SARS.

Results

SARS data was collected on 238 mission-ready F-15C pilots from 11 squadrons stationed at 4 different Air Force bases. Two hundred and six of these pilots provided written definitions of SA. The first column in Table 6.2 lists the seven phrases most frequently used in defining SA. The second column shows the seven most highly rated elements of SA. There is considerable agreement between the phrases used to define SA and the element ratings. In addition, both the phrases and the element ratings indi-

TABLE 6.2
Seven Phases Most Commonly Used by Pilots to Define SA and the Seven Most Highly Rated Elements of SA Listed in Decreasing Frequency of Occurrence

Most Commonly Used Phases to Define SA	*Most Highly Rated Elements for SA*
Composite 3-D image of entire situation	Use of communication information
Assimilation of information from multiple sources	Information integration from multiple sources
Knowledge of spatial position or geometric relationships among tactical entities	Time-sharing ability
Periodic mental update of dynamic situation	Maintaining track of bogies and friendlies
Prioritization of information and actions	Adjusting plan on-the-fly
Decision making quality	Spatial ability to mentally picture engagement
Projection of situation in time	Lookout for threats from visual, radar warning receiver, radar

TABLE 6.3
SARS Intercorrelations

	1	2	3	4	5
1. Supervisor SARS					
2. Peer—Fighter pilot ability	0.89				
3. Peer—SA ability	0.91	0.98			
4. Peer—Rank order	0.92	0.91	0.92		
5. Self-report SARS	0.45	0.56	0.57	0.49	

cate that a significant component of SA involves assimilating and using information to guide action.

Analyses of the peer and supervisory SARS indicated that the pilots reliably classified their fellow pilots in terms of SA. Internal consistency was computed for all 31 items on the supervisory SARS. The resulting measure, Cronbach's coefficient alpha, was 0.99. Interrater reliability was also estimated for the supervisor and peer SARS using an analysis of variance procedure (Guilford, 1954). For the supervisor SARS, these analyses indicated that the average reliability of each supervisor's ratings was 0.50 and the average reliability of the pooled supervisor ratings was 0.88. Similarly, the peer SARS showed an individual reliability of 0.60 and a combined reliability of 0.97. Additional detail concerning the analyses of the SARS data is found in Waag and Houck (1994).

As shown in Table 6.3, there was substantial agreement between supervisor and peer SARS. Table 6.3 also indicates that there is noticeably less agreement between the self-report SARS and the other SARS.

MEASURING SA IN SIMULATED AIR COMBAT MISSIONS

Although the SARS data indicate high reliability and consistency between raters, they do not empirically relate judged SA to pilot performance in air combat missions. In an attempt to determine the relation between SA and mission performance, a composite SA score was developed for each of the 238 pilots. These scores, based on the peer and supervisory SARS, were scaled with a mean of 100 and a standard deviation of 20. Based on this composite score, a sample of 40 mission-ready flight leads was selected to fly a series of multiship air-to-air combat simulations. Mission qualification level was held constant for the simulation portion of this effort because current flight qualification was highly correlated ($r = 0.82$) with the composite SA score (Waag & Houck, 1994). The mean SA scores for the 40 flight leads were 106.3 with a standard deviation of 17.4. An additional 23 mission-ready pilots flew as the flight lead's wing during the course of the simulation.

The combat simulations were flown in the Air Force Research Laboratory's multiship training research facility in Mesa, Arizona. This facility provided the flight simulators and associated simulations necessary to allow the participants to fly realistic combat missions in multibogey, high threat scenarios. The two F-15C pilots flew high fidelity F-15C simulators operating on a secure simulation network. The simulation network also included other manned and computer-controlled aircraft, computer-controlled surface-to-air threats, and a manned air weapons controlled simulator. Additional details concerning the simulation are found in Waag, Houck, Greschke, and Raspotnik (1995).

Each flight lead flew nine simulator sorties over 5 consecutive days. During each sortie four separate engagements were flown. Each engagement consisted of a different scenario representing the same basic mission. After each engagement, the simulation was reinitialized to the appropriate starting conditions and the new scenario began. Scenarios increased in complexity throughout the week.

Scenario Design

Figure 6.1 illustrates a moderately difficult scenario. In this defensive counterair mission, the two F-15s defended an airfield. The attackers consisted of two bombers escorted by two fighters. The scenario began with the enemy aircraft 80 nautical miles (nm) away from the airfield. The enemy fighters were flying at 20,000 ft and the bombers at 10,000 ft. There was a lateral separation of 10 nm between the fighters and the bombers. At 35 nm, the fighters maneuvered rapidly and descended to 3,500 ft. At 15

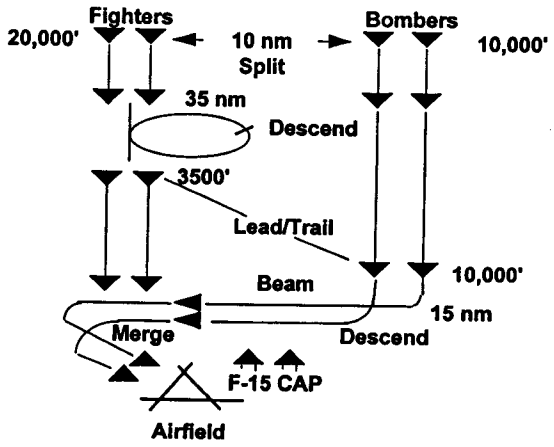

FIG. 6.1. Typical defensive counterair scenario used as part of the combat mission simulations.

nm, the bombers performed a hard right turn and descended to 2,500 ft. The purpose of these maneuvers was to momentarily break the F-15s' radar contact and disrupt the F-15 pilots' ability to identify, target, or engage the enemy aircraft.

Rating Mission Performance

The basic approach taken toward SA measurement was through scenario manipulation and performance observation as suggested by Tenney et. al., (1992). Other approaches, such as explicit probes and the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1995b), were considered. These other approaches were rejected because of a need for measures that could be used during operational training either in simulators or actual aircraft. All of the scenarios contained enemy maneuvers designed to "trigger" information seeking and decision making by the flight lead. In essence, these trigger events serve as SA probes in a naturalistic environment.

As Kelly (1988) pointed out, measuring air combat skills presents a number of challenges. The fluid, dynamic nature of air combat, combined with the number of alternative tactics and techniques available to the pilot, make objective performance measurement extremely difficult. Even when objective data is available, it is often difficult to interpret the importance of that data. Because of the difficulties involved in establishing the relation between objective measures such as radar locks or engagement parameters and SA, it was decided to rely on behavioral observation by SMEs who were unaware of the SA scores of the pilots they were observing. Two SMEs, retired fighter pilots with extensive experience in air combat and training, watched each engagement in real time and independently completed an observational checklist. To assist them in evaluating pilot performance, cockpit instruments, intraflight communications, and a plan view display of the engagement were available throughout the engagement. After each simulator session, the two SMEs discussed each engagement and completed a consensus performance rating scale containing 24 behavioral indicators based on the SARS. These 24 behavioral indicators were the elements of SA listed in Table 6.1, less the 7 general traits of SA. In addition, the SMEs also wrote a critical event analysis for each mission that identified events that were critical to the outcome of the mission and indicative of the pilot's SA.

Results

Figure 6.2 shows the composite SA scores obtained from the SARS and the mean SA score assigned by the SMEs based on their observation of each pilot's performance during the simulated air combat missions. The Pear-

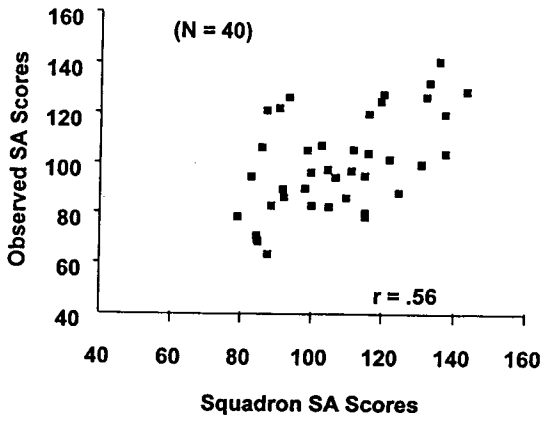

FIG. 6.2. Mean SA score for combat mission simulation as a function of squadron SA score.

son product moment correlation between these scores is 0.56 ($p < .001$, $df = 38$) and indicated a significant relation between squadron ratings of SA and performance in these simulated air combat missions.

SA Ratings and Specific Behaviors

Having established that squadron SA ratings and simulator performance ratings are related, the next question is whether specific behaviors associated with rated SA could be identified. Two aspects of pilot behavior were examined: errors (as identified by SMEs) and communication patterns.

Identification of Errors. Two SMEs reviewed tapes of selected missions and identified errors made by pilots while flying complex scenarios. Four engagements for each of 8 pilots were chosen for this analysis. The pilots selected were the top four and bottom four individuals based on squadron SA rating. The SMEs independently reviewed each of the resulting 32 engagements. They were instructed to identify and record any pilot action or inaction that they considered an error. No strict definition of "error" was provided. The idea was to allow the SMEs to identify without constraint any action or inaction they felt was inappropriate given the situation and the available data. After identifying errors separately, the SMEs reviewed the scenarios together and agreed upon a final set of errors. They also agreed on a list of mutually exclusive error categories and placed each identified error in a category. These categories were later grouped into two more general classes: "decision errors" and "information acquisition errors."

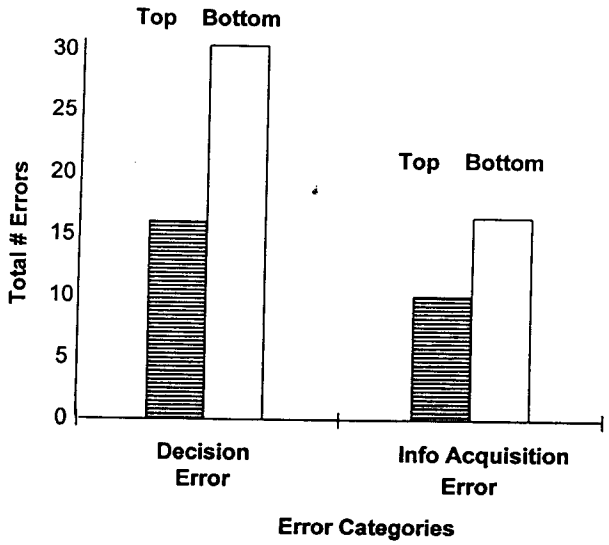

FIG. 6.3. Total number of decision and information errors made during four simulated combat engagements as a function of squadron SA score.

Figure 6.3 shows the number of decision errors and information acquisition errors made by top- and bottom-ranked pilots for the squadron SA ratings. Although the number of errors in each category is small, a relatively large number of decision errors were observed among pilots who were ranked lowest in SA by their supervisors and peers at the squadron. Figure 6.4 shows the breakout of decision errors into more specific categories. Inspection of Figure 6.4 shows that the difference between low-SA and high-SA pilots in the number of decision errors is largely accounted for by errors in tactics selection and flight leadership. Thus, subjective ratings of SA by peers and supervisors are related to more than just performance ratings made by other SMEs; they are also related to observable behaviors in complex combat scenarios.

Communication Patterns. Communication is an important source of information to support SA in air combat pilots. Therefore, some aspects of communication behavior may be related to squadron-rated SA. The nature of such a relation is, however, hard to predict. For example, consider the frequency of calls to a wingman or air weapons controller (AWC) to request information. One view is that high-SA pilots will request information less frequently than low-SA pilots because their SA is already good. The opposite view is that high-SA pilots maintain their SA advantage, in part, because they request information more frequently. To address such issues, communications from the top four and bottom four pilots in squadron SA rating

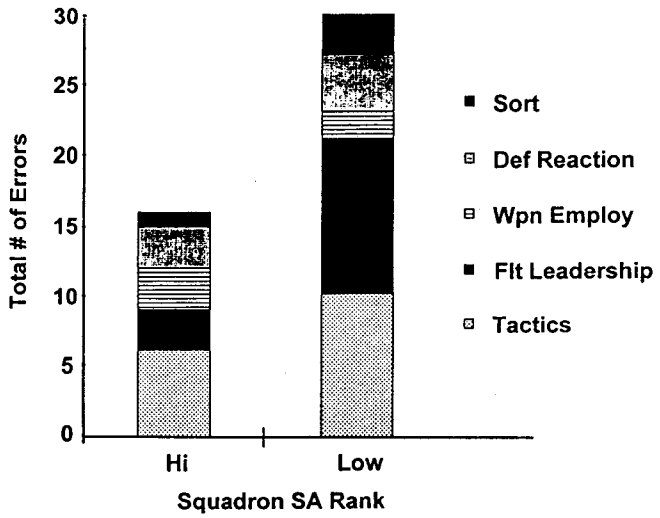

FIG. 6.4. Types of decisional errors made as a function of squadron SA score.

were transcribed. Three engagements were analyzed for each pilot. Each call made during each of the engagements was transcribed and categorized according to the initiator of the call (lead, wing, AWC); the intended recipient (lead, wing, AWC, entire team); and the purpose (providing information, acknowledging, directing, requesting, informing, or uncodable).

Because engagements differed in length, comparisons of communication frequency were generally made in terms of the number of calls per minute, or call rate. As shown in Fig. 6.5, call rates were higher for high-SA leads than for low-SA leads for every kind of call except acknowledgments. High-SA leads both provided and requested information more frequently than did low-SA leads. High-SA leads also directed other team members more frequently. Like the error data, the communication data analyzed so far suggested that subjective ratings of SA are related to behavioral differences during simulated combat scenarios.

DISCUSSION

These initial results in developing measures of SA that can be used in a squadron's operational training environment are encouraging. These results indicated that SA is a construct that has meaning and can be used by both peers and supervisors to classify mission-ready pilots. They also indicated that squadron ratings of SA are related to relevant behaviors (e.g.,

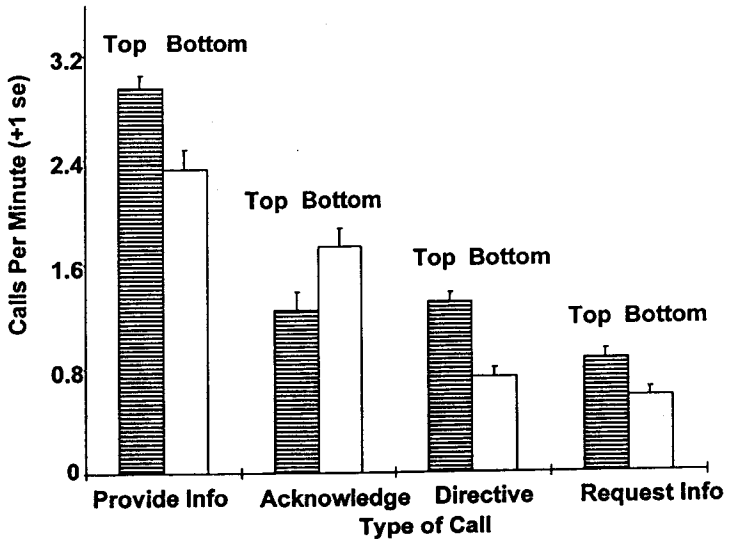

FIG. 6.5. Number of radio calls per minute as a function of squadron SA score and type of call.

errors, communications) and are correlated with mission success in simulated air combat missions.

Although this approach to SA measurement may be classified as subjective rather than objective, this is an oversimplification. All measurement approaches ultimately involve assigning numbers to events according to an explicit set of rules (Stevens, 1951). The distinction between objective and subjective measures simply indicates whether a human observer is an integral component of the measurement instrument. Objective measurement involves data that is generated independently of the human observer. Ideally, this data is generated, recorded, and scored without the intervention of a human observer. Subjective measurement, on the other hand, requires human observers to generate the data itself. Although Muckler (1977) argued that there is no such thing as objective measurement in the strict sense, the distinction continues to be made and objective measures are often preferred to subjective measures. The reason for this preference is that subjective measures are frequently seen as being contaminated by the human observers during the act of measurement. Because objective meas-0ures, on the other hand, are relatively independent of human observers, they are seen as "truer" measures of the construct under study.

Unfortunately, objective measures often fail to capture the richness and complexity of human performance (Kelly, 1988; Meister, 1989; Vreuls & Obermayer, 1985). One reason for this is that objective measures are essentially reductionistic and are therefore best suited for recording the fun-

damental dimensions of performance (e.g., latency, amount, and deviation). Although these fundamental measures provide us with data that is less subject to error, they also frequently fail to provide us with information concerning the contextual nature of skilled performance. Subjective measures seem more closely related to higher order psychological constructs. The data they produce appears to reflect a synthesis of the more molecular behaviors and to reflect more global dimensions such as interpreting, judging, and deciding—the very essence of SA.

Potential Problems With Subjective SA Ratings

Subjective ratings are a useful tool because they capture the complexity of SA. Unfortunately, various kinds of problems can limit their usefulness. At the heart of many of these problems is the definition of SA itself. Disagreement about exactly what constitutes SA—between raters and other raters, raters and scientists, or even within an individual rater at different times—will tend to yield rating results that are difficult to interpret.

A key definitional issue is the distinction between momentary SA and SA ability. Momentary SA is demonstrated when one interrupts a mission at a particular point and tests current-situation knowledge (cf., Endsley, 1995a). SA ability is the tendency to maintain good SA in a variety of situations. The Air Staff definition of SA used throughout this effort includes elements of both.

There are serious interpretational problems associated with the assessment of both SA ability and momentary SA. Subjective-ability-measurement problems are illustrated by the task of obtaining ratings of SA ability from squadron supervisors and peers. A primary problem is to communicate to the raters the difference between SA ability and overall piloting skill because several important aspects of pilot skill do not fall within the definition of SA. This problem was resolved by requiring separate ratings of SA and "fighter pilot ability."

Another inherent problem is that raters may be unduly influenced by a pilot's credentials. Rank, hours of experience, qualification status, and participation in special exercises could influence rated SA. Even if the rater is explicitly instructed to disregard credentials in producing a rating, difficulty can be expected because the rater's interpretation of past experiences with the pilot, upon which the SA rating should be based, is colored by the pilot's achievements. The obvious approach to mitigate this problem is to collect data on credentials and experiential factors so that the differential utility of the SA ratings over and above these factors can be assessed. In this study, hours of experience in the F-15 was correlated 0.62 with composite SA rating. When four other experience factors (qualification, Fighter Weapons School graduation, and participation in two exer-

cises) were used to predict composite SA rating, the multiple correlation was 0.85 (Waag & Houck, 1994).

These problems—differentiating SA from performance and credentials—also appear in attempts to assess momentary SA. The SMEs' ratings of SA during each combat scenario can be viewed as a time-averaged momentary-SA measure. SMEs were instructed to rate the pilot's SA during a particular scenario rather than try to assess the pilot's overall SA ability. These instructions were intended to induce the raters to attend to the pilot's SA at various moments during the engagement and arrive at a composite of these momentary SA assessments, as opposed to a general prediction of the pilot's SA ability in other situations. Interviews with the raters, during which they were asked to explain their ratings, suggested that they in fact evaluated SA as it was demonstrated during the engagement rather than predicting overall SA ability. This conclusion is supported by rater comments such as "I can tell that this pilot is very good, but this time he was doing X and didn't notice threat Y, so we had to give him a low rating."

As noted earlier, SMEs were not informed of the pilot's level of experience, most of the pilots were of the same rank, and all of them were at the same qualification level (4-ship flight lead). Although this may have mitigated any tendency for raters to base ratings on pilots credentials, it did not solve the problem of differentiating SA from performance in a scenario. Even though the distinctions between SA, performance, and mission outcome were stressed in instructions to the SMEs, statements by the SMEs in the mission summaries they provided after each scenario suggested that some confounding occurred. In particular, mission success or failure undoubtedly had some influence on SA ratings.

The difficulty of rating SA independently of pilot performance is exacerbated by the inclusion of performance-like elements in the definition of SA. For example, the Air Staff definition includes ". . . the ability to . . . execute tasks . . ." (Carroll, 1992, p. 5). This is more than an afterthought—it is a critical feature of the definition for some investigators. For example, Vidulich (1992) wrote "good SA implies a capability to respond appropriately, not just the possession of an accurate assessment" (p. 17). Wickens (1995) also argued that definitions of SA are incomplete without some reference to "the capacity to respond appropriately to events" (p. K2–2). On the other hand, Endsley (1995a) stressed the importance of separating a full assessment and comprehension of the situation from the decisions and responses based on that comprehension. At an academic level, these may not be substantive disagreements. Most investigators seem to agree that while the knowledge of which responses are appropriate to a situation is part of SA, the responses themselves are not. However, to the rater the fact that a pilot responded appropriately is usually good evidence that knowl-

edge of appropriate action was present. Thus, to the extent that knowledge of appropriate courses of action is stressed in the definition of SA, it will be difficult to rate SA independently of performance.

There is another way of defining SA that may be more useful for guiding the process of obtaining accurate ratings. This approach recognizes that all aspects of momentary SA are eventually reducible to some form of knowledge or information in working memory. At any given moment, the information in a pilot's working memory can come from immediate perception (Endsley's Level 1 SA); associations from long-term memory, including retrieved interpretations of perceived information (Endsley's Level 2 SA); and inferences generated by combining and manipulating information in working memory, including predictions about future events (Endsley's Level 3 SA). However, not all working memory information is SA-related (e.g., daydreams are probably not). Therefore, SA could be defined as knowledge (in working memory) about elements of the environment. For a pilot "knowledge about elements of the environment" must be interpreted broadly to include not only information about the location, capabilities, and intentions of threats, but also such things as offensive and defensive status and likely future events. A potential advantage of explicitly focusing on the information that makes up SA is that raters can be instructed to look for evidence that a pilot has various pieces of knowledge.

Of course, not all knowledge is equal in its impact on mission success. Even if the information in a pilot's working memory could be measured objectively, there remains the problem of assigning a weight to each piece of information according to its relevance. These weights would no doubt be constantly changing as the mission evolves. Any objective measurement technique based on assessing specific pieces of knowledge must somehow cope with this daunting weight-assignment problem.

In contrast, a human rater "automatically" takes into account the importance of various components of SA in generating a rating. This should be one of the biggest advantages of using subjective measures; however, this advantage only applies if the rater focuses on SA as knowledge and does not confound SA with performance or mission outcome.

Generalizability to Other Domains

This research obtained SA ratings of operational pilots from peers and supervisors with minimal disruption of their squadron operations and demonstrated that these ratings were predictive of SA ratings independently obtained during simulated combat missions. It represented a relatively efficient method of obtaining SA ratings in both field and laboratory environments and demonstrated that SMEs are capable of moderate to good agreement regarding an individual's SA.

Although this research focused on the SA of mission-ready F-15C pilots, global SA ratings obtained from peers or supervisors, momentary SA ratings based on direct observation of job performance, or both can be applied to measuring SA in other domains such as commercial aviation, medicine, and command and control. Because the specific details of the research methodology will vary depending upon the domain being studied, we believe anyone attempting to apply this approach should adhere to the following guidelines:

1. Regardless of whether global or momentary SA ratings are used, it is essential to develop a set of behavioral indicators that will provide raters a common framework for judging SA.
2. Field raters providing global SA ratings must have had multiple opportunities to observe the ratees perform in their natural environment and these raters must be given the opportunity to omit rating individuals if they feel they have not had sufficient opportunities to observe that individual.
3. Interrater reliability of global SA ratings improves as the number of raters increases.
4. Job samples used to provide the opportunity for momentary SA ratings must provide the opportunity for the individual to exhibit a variety of behaviors that are indicative of SA.

Utility of the SA Concept

The potential for confounding SA with performance strikes at the heart of the utility of the concept. SA is a useful concept to the extent that it suggests measures that add to our ability to predict which training techniques, pilots, and displays will lead to the best mission outcomes. A subjective SA rating can be used as an independent, dependent, or intervening variable. As an independent variable, it could help predict which pilots will achieve success in combat. As a dependent variable, it could provide a sensitive method for evaluating the impact of new displays. As an intervening variable, it could help explain why different training techniques lead to different mission outcomes.

Unfortunately, if raters use mission outcomes to infer SA, little is gained from introducing the SA variable; it will have little or no differential utility. Our experience suggests that, to obtain useful SA ratings, raters must focus on what a pilot probably knows about the elements of a situation. As discussed previously, this can include valid inferences, predictions, or both about present or future events within the scenario. We used various tools to help raters read the pilot's mind. A video record of the

radar display is particularly important. We also used knowledge-based checklists, and, for some scenarios, the pilot's eye movements were recorded and the raters were provided a video showing the pilot's point of gaze. Of course, raters will still make inferences from the actions taken by the pilot. The rater, however, must be constantly aware that executing a maneuver well or making a correct tactical decision is not necessarily proof of good SA. If the rater can maintain this mental set during the rating process, then it is possible to use a pilot's behavior to infer at least some of the knowledge leading to that behavior.

CONCLUSION

Obviously both subjective and objective measurement approaches are necessary to develop an understanding of SA. Objective measures are important because they provide a necessary check on subjective judgments. Suppose a pilot with impressive credentials misses a critical radar sampling assignment, then pulls the mission out of the fire with superior stick-and-rudder skills. Behavior-based, face-valid indicators of SA would help assess the extent to which ratings in such situations are driven by mission outcome rather than SA. Subjective measures, on the other hand, help to assure that critical aspects of SA are actually being assessed. For example, there are several ways that a pilot can be aware of the location of a particular enemy aircraft (e.g., radar, wing, AWC). An objective measure of the number of enemies sampled on the radar might well be a misleading and incomplete indicator of SA if important information on enemy locations was given by a weapons director and only the critical targets were sampled on radar. Objective measurement technology will doubtless move toward configural assessment of some aspects of SA; however, it will be difficult to mimic the inferences that a human SME makes to assess the cognitive aspects of SA that are not directly measurable.

We believe that the critical SA measurement issues concern how the definition of SA is refined, and which measurements provide the best information for designing and evaluating aircrew training. Regarding the definition of SA, we have argued that the development of both objective and subjective measures will be facilitated by defining momentary SA in terms of specific kinds of knowledge, where knowledge is broadly interpreted to include the pilot's inferences about the enemy's status and intentions, and guesses about likely future events. Viewing SA in terms of constituent knowledge should also facilitate the design and evaluation of aircrew training. It is important, however, not to oversimplify this knowledge-based view of SA. As is always the case in training, simply providing important information is no guarantee that the information will be in the

pilot's working memory at the time it is needed. Thus, it is also necessary to focus research and measurement development on the processes by which knowledge is acquired during flight.

ACKNOWLEDGMENTS

The authors wish to thank Don Vreuls for his helpful comments on an early version of this chapter and Wayne Waag for his role in establishing the Air Force Research Laboratory's situational awareness research program. The authors also wish to thank Bart Raspotnik, Brian Schreiber, and Kelly Lee for their role in analyzing the communication and error data. Portions of this chapter were presented at the International Conference on the Experimental Analysis and Measurement of Situation Awareness, Daytona Beach, FL, 1–3 November 1995.

REFERENCES

Carroll, L. A. (1992). Desperately seeking SA. *TAC Attack (TAC SP 127–1), 32,* 5–6.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65–84.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1993). *An information processing classification of beyond-visual-range air intercepts* (AL/HR–TR–1993–0061, AD A266 927). Williams Air Force Base, AZ: Armstrong Laboratory.

Kelly, M. J. (1988). Performance measurement during simulated air-to-air combat. *Human Factors, 30,* 495–506.

Meister, D. (1989). *Conceptual aspects of human factors.* Baltimore, MD: Johns Hopkins University Press.

Muckler, F. A. (1977). Selecting performance measures: "Objective" versus "subjective" measurement. In L. T. Pope and D. Meister (Eds.), *Symposium proceedings: Productivity enhancement: Personnel performance assessment in Navy systems* (pp. 169–178). San Diego, CA: Navy Personnel Research and Development Center.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1,* 45–57.

Stevens, S. (1951). Mathematics, measurement, and psychophysics. In S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Tenney, Y. J., Adams, M. J., Pew, R. W., Huggins, A. W. F., & Rogers, W. H. (1992, July). *A principled approach to the measurement of situation awareness in commercial aviation* (NASA Contractor Report 4451). Langley, VA: National Aeronautics and Space Administration.

Vidulich, M. (1994). Cognitive and performance components of situation awareness: SAINT team task one report. In M. Vidulich, C. Dominguez, E. Vogel, & G. McMillan (Eds.), *Situation Awareness: Papers and Annotated Bibliography* (AL/CF–TR–1994–0085; pp. 17–28). Wright-Patterson Air Force Base, OH: Armstrong Laboratory.

Vreuls, D., & Obermayer, R. W. (1985). Human-system performance measurement in training simulators. *Human Factors, 27,* 241–250.

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine, 65*(5, Suppl.), A13–A19.

Waag, W. L., Houck, M. R., Greschke, D. A., & Raspotnik, W. B. (1995). Use of multiship simulation as a tool for measuring and training situation awareness. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD–CP–575; pp. 20-1– 20-8). Neuilly Sur Seine, France: AGARD.

Wickens, C. D. (1995). Situation awareness: Impact of automation and display technology. In *Situation Awareness: Limitations and Enhancement in the Aviation Environment* (AGARD–CP–575; pp. K2-1–K2-13). Neuilly Sur Seine, France: AGARD.